Modeling Electoral Outcomes: A Logistic Regression Analysis of County-Level Voting Trends in Virginia

DS 3001 – Foundations of Machine Learning

Caroline Sullivan, Joe Thompson, Sasha Porter, Genevieve Purcell

ABSTRACT:

This study examines how historical voting trends can forecast the outcomes of presidential elections using Virginia's county-level data from 2000 to 2020. By employing supervised learning with binary logistic regression, the research explores the relationship between voting patterns and party outcomes in Virginia. The model uses features such as party vote share trends, voter turnout changes, and margins of victory to classify county-level election results as either Democratic or Republican wins. Our research question investigates the extent to which historical voting trends and past election results can accurately predict future county-level election outcomes. Key findings include that Republican vote share trends have the largest impact on predictions, with a significant upward trend strongly correlating to Democratic losses. Interestingly, an increase in Democratic vote share does not necessarily translate to a Democratic win in historically Republican-dominated counties. These patterns highlight the complexity of regional voting behavior and the influence of historical party dominance. The model achieved a 78% accuracy rate in predicting county outcomes for the 2024 presidential election, performing well in forecasting Republican wins but showing limitations in capturing Democratic victories, particularly in areas with shifting demographics. Challenges include accounting for external variables, such as political and economic shifts, and the lack of county population data to

evaluate the impact of population distribution on predictions. Future research should incorporate demographic factors, population density, and alternative classification techniques to refine predictive accuracy. This study demonstrates the utility of historical voting data for electoral forecasting and underscores the importance of regional and temporal trends in understanding political outcomes. By comparing model predictions to 2024 results, this research provides insights for electoral strategists and contributes to the growing field of predictive modeling in political science.

INTRODUCTION:

Predicting presidential election outcomes has fascinated data analysts, policymakers, and political scientists for decades. Understanding these election outcomes has become an essential component of political science, particularly in swing states like Virginia. Following the 2024 presidential election, we combined our interest in the election with our knowledge of data science and machine learning to forecast the election results in Virginia.

This paper investigates the predictive power of historical voting trends at the county level in Virginia to forecast the results of the 2024 U.S. presidential election with our research question asking: *How accurately can previous election results and historical voting trends predict the outcomes of future elections?* This study uses binary logistic regression to analyze data from six past presidential elections, from the years 2000 to 2020, and identify patterns in voting behavior linked to party outcomes.

The analysis focuses on key variables, including the trends in party vote share, voter turnout changes, and margins of victory, to classify each county's election results as a Democratic or Republican win. Based on our results, there is value in using historical voting

trends to predict election outcomes, but there are also limitations as well. The model performed well in predicting Republican wins, but was less accurate with Democratic outcomes. Our results show that Republican voting trends strongly impact election predictions, despite the increasing Democratic support in Virginia. This reflects complex voting patterns and lasting party influence in the US.

When comparing our model's results to the actual results of the 2024 presidential election, we concluded that the model achieved a strong 78% accuracy in predicting results. Despite this, it tended to underestimate Democratic wins. The findings suggest that while historical voting trends are useful, they are not direct indicators of future results. External variables must also be considered to improve the accuracy of future predictions.

Accounting for changing voter behavior and the impact of external factors such as demographic shifts, economic conditions, and political events is a key challenge faced in this report. For instance, the COVID-19 pandemic in 2020 led to significant changes in voting methods, including an increase in absentee ballots. This may have altered traditional voting patterns. Similarly, shifts in population, such as growing suburban areas and increasing racial and ethnic diversity, can influence voting trends and party alignment. These external factors add complexity to the prediction model and may not be fully captured in historical voting data alone.

Our research and model is significant because it shows how using historical voting data can help predict election outcomes, while also highlighting its limitations. By comparing predictions to the 2024 election results, it helps political strategists understand how past trends influence future votes. This research adds to the growing field of electoral modeling and encourages future studies to consider more factors for better predictions.

In the future, we can improve our models by including external factors that affect elections, like changes in voter sentiment, demographics, and political events. As we refine our models, using real-time data will be important to keep forecasts up-to-date with political shifts. In the following sections of this paper, we will delve deeper into the methodologies employed in our analysis, the limitations of our model, and explore potential adjustments that could enhance its predictive power.

DATA:

Our dataset contains voting information for Virginia's presidential elections from 2000 to 2020. It includes the election year, the state (which is exclusively Virginia), and detailed data for each county, including their FIPS code (a county-specific identifier), the number of votes cast for each presidential candidate, and the total votes in each county. The candidates are categorized by political party—Democrat, Republican, and Other—providing a clear breakdown of the political landscape during each election.

Historical voting data is often a reliable predictor of future elections, and this dataset will be crucial for forecasting the 2024 presidential outcome in Virginia, both at the county level and for the state as a whole. By providing detailed information on the distribution of votes among candidates and parties from previous elections, the data allows for a granular analysis. With the results broken down by county, we can examine voting patterns across regions, identifying trends like consistent party leanings or significant shifts in voter behavior across election cycles. Leveraging these patterns from prior years will enable us to make more accurate predictions for 2024. Additionally, by analyzing total vote counts over time, we can gain valuable insights into voter turnout trends, helping to refine our predictions even further.

In 2020, we observed a significant change in how the county's presidential votes were recorded. Unlike previous years, where votes were categorized as "TOTAL" votes, the data was segmented into absentee, election day, and provisional votes, rather than presented as a single total. We believe this shift was likely influenced by COVID-19, which limited in-person voting options. To ensure consistency across all years, we will likely need to use Python to aggregate these different vote types for 2020, so that each year's data aligns with a uniform format.

A challenge we encountered is that for the year 2000, the Green Party is listed separately, alongside the Democratic, Republican, and Other categories. Similarly, for 2020, the Libertarian Party appears as a distinct category. To maintain consistency in formatting across all years, we recommend consolidating both the Green and Libertarian parties under the "Other" category. This will ensure uniformity in party classification throughout the dataset.

METHODS:

Our study's research question is: *How accurately can previous election results and historical voting trends predict the outcomes of future elections?* In this study, each observation will be the number of votes a candidate received by party in each presidential election, held every four years, for each county in the state of Virginia. Specifically, we will be looking at whether they win or lose. Since we are using historical data to predict labeled outcomes, we will be conducting a supervised learning study. Our goal is to predict a binary outcome—win or loss—for each candidate or party, making this a classification analysis. Specifically, we aim to determine whether the winning party in previous elections can help forecast performance in the 2024 election.

To analyze this we will use logistic regression to classify outcomes by Democratic win or Republican win for each county and presidential election. Specifically we will use binary logistic regression because there are a finite number of outcomes: 1 representing a Democratic win and 0 representing a Republican win. We focused on the Democratic and Republican parties because the outcomes in each county in Virginia for each election we have data for consisted of only these two parties, with them gaining a large majority of the votes. First, we will transform the dataset so that a record will represent the data pertaining to a single county for one particular presidential election; there will be one row for each county and election pair. By training the model on data points from past elections up until the most recent election, we'll test how well it predicts the upcoming 2024 election outcome in each county. We will know if our approach "works" if we can effectively predict the results of the next election based on previous election data. Success is defined by how accurately our model predicts the 2024 election results. We will then compare the predicted outcomes to the actual outcomes of the 2024 presidential election for each county once the results come out.

Challenges could arise from changes in voting behavior, such as increased mail-in voting since 2020, and external influences, such as economic or political shifts. These challenges may not be captured in the data, so we may not be able to directly address these weaknesses. We will still remain aware of their potential impact on our results. If our approach fails, we will learn how to adjust our analysis and fix any easily identifiable problem. To give the logistic regression model variables based on historical voting trends to learn patterns from, we will engineer variables that represent the vote shares for each party in each county, as well as its trend from the previous election. This will be done by calculating the percentage of votes each party received in each county per presidential election year and then taking the percentage difference of the vote

share from the previous presidential election. Additionally, we will look at the margin of victory between the two leading parties and how the total number of votes, or voter turnout, changes throughout the years. We could also sum or average votes across other features besides by party to develop a more well rounded illustration of the data and what the data show.

We will communicate our results using a combination of tables and visualizations. This will include maps of Virginia, showing the winning party by county in current and historical contexts. Graphs and visualizations highlighting the historical trend of the winning party by county and any correlations between county demographics and party leaning will be created. A table of regression coefficients will provide insight into how past presidential election data contributes to predicting the current year while controlling for other factors.

RESULTS:

2024 is a historical year with significant implications of the presidential election. Given that it is the season for a pivotal presidential election, we wanted to use machine learning tactics to predict the election outcome for the state of Virginia. Our prediction question is as follows:

Considering voting data from the 2000, 2004, 2008, 2012, 2016, and 2020 presidential elections, how significant are historical voting trends in accurately predicting presidential election outcomes for specific counties in the state of Virginia?

We found that, by using historical voting trends such as the trend of Democratic and Republican vote share, our model predicted the results of the 2024 Presidential Election for each county in Virginia with 78% accuracy.

In this discussion, we will highlight historical voting trends for the state of Virginia in parts I and II. Then we will discuss the results we derived from our logistic regression model in part III. To conclude, in part IV we will compare our findings and predictions to the actual results of the 2024 presidential election for the state of Virginia.

I. Historical Voting Trends

Figure 1:

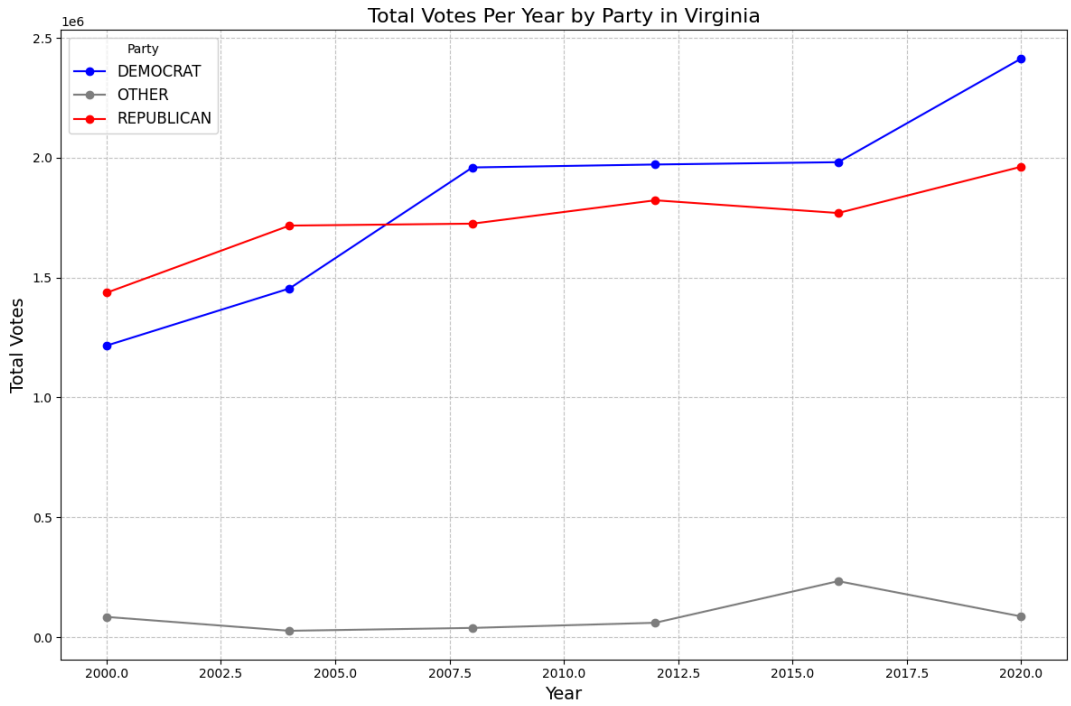

Figure 1 illustrates voter turnout trends over the past 20 years. The blue line represents the number of votes received by the Democratic party each year, while the red line represents the number of votes received by the Republican party each year. The grey line aggregates votes for third-party candidates not affiliated with a major party.

The data reveals a significant upward trend in Democratic voter turnout with a notable spike in Democratic votes in 2020, shown as the blue line. This is likely due to higher

Democratic engagement and possibly demographic shifts favoring the party. In contrast, the red line is showing the Republican votes and how they had moderate growth from 2000 to 2004, but since then have stagnated with only slight fluctuations. The grey line shows that third-party votes have had minimal impact overall, with a slight increase in 2016 which likely represents a third-party candidate gaining traction in a specific election.

Overall, the total number of votes has increased steadily, reflecting growing voter participation in recent elections. This trend may be driven by heightened national political polarization, which tends to boost voter engagement. Contributing factors include population growth and demographic changes, particularly in Virginia. Over the past few decades, the state has seen significant population increases and greater diversity, especially in urban and suburban areas, which traditionally lean Democratic.

II. Past Results

Figure 2:

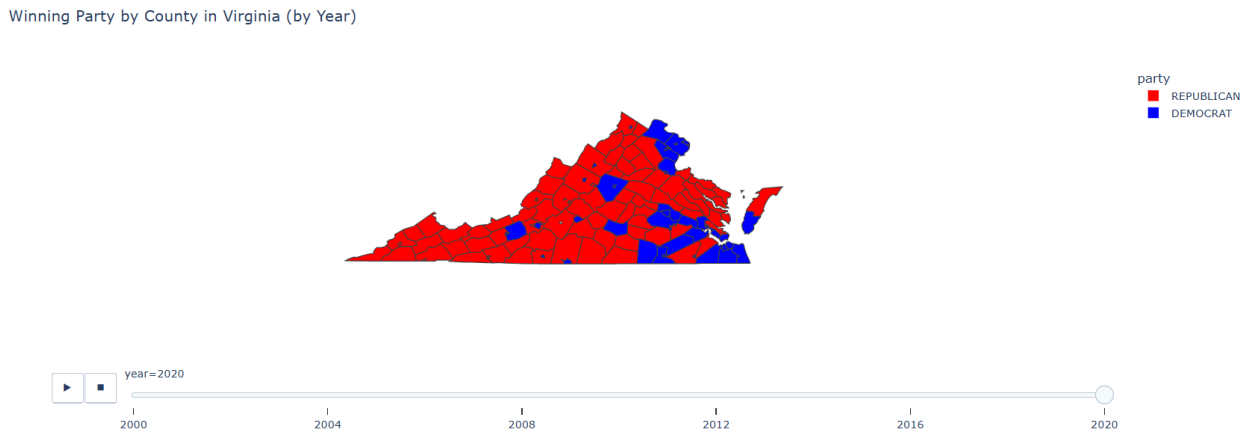

Figure 2 is a snip from a video of an interactive map of Virginia showing the difference in county voting results over the 20 year period. Counties are colored by who won their majority vote, with red being Republican and blue being Democratic.

There are many counties who have stayed the same color (or same majority vote) over the past 20 years, but there are many who have changed their color. In 2000, there were a lot more red counties than in 2020, and this is likely due to the change in Virginia's population and move to more urban and suburban areas. Election-specific dynamics is also a big factor in what the majority vote is in each county as many people may be partial to a specific candidate.

III. Model Results

Based on our logistic regression model, we ran a regression analysis to find the following coefficient results:

Figure 3: Coefficient Results

Feature Variable	**Coefficient**
democratic_share_trend	-0.357703
republican_share_trend	-3.409907
other_share_trend	-0.290734
turnout_change	0.241128
normalized_margin_of_victory	-0.178615

To predict the election outcomes in Virginia counties, we derived feature variables based on the historical voting trends found in Virginia's voting data. These variables consisted of vote share trends by party (Democratic, Republican, and other), the trend of voter turnout, and the margin of victory. With the use of these feature variables, our model resulted in a 77% training accuracy and 75% testing accuracy. When running the regression on testing data, the model was found to have a 77% precision when classifying a record as 0 and a 62% precision when classifying it as 1. This indicates that it was more accurate in predicting a Republican win over a Democratic win. Furthermore, there were 32 false negatives (Democratic wins misclassified as Republican) compared to 9 false positives (Republican wins misclassified as Democratic), suggesting the model more frequently underestimated Democratic victories. The model's higher accuracy in predicting Republican outcomes can be explained in the coefficients for our feature variables. The results suggest that the Republican vote share trend impacts the accuracy of our model the most, with a coefficient of -3.4. Thus, a significant upward trend in the percentage of votes a Republican candidate receives is a strong predictor of the likelihood of a Democratic loss. We can also see that the Democratic vote share trend has a coefficient of -0.36, implying that a rising trend in Democratic vote share is less likely to have a Democratic candidate win. This might seem counterintuitive, but it is important to note that an increasing Democratic trend does not directly imply a Democratic win in that county. Because these variables have large negative coefficients, our model leans towards a lower likelihood of a Democratic win.

IV. Comparison to Actual Results

To compare our predictions to the actual results of the 2024 presidential election, we estimated the Democratic, Republican, and other vote counts through the use of the vote count trend for the respective party. We then derived the rest of the feature variables from these

estimates and the data from the 2020 election. The model had a 78% accuracy when predicting the results of the voting data of the 2024 presidential election based on trends from prior elections. One reason for this discrepancy is that voters' tendencies can change as a result of external variables, suggesting that they can flip between years. Some issues can appear more prevalent and be more likely to influence voting decisions between election cycles that our model can not account for. The model's performance highlights its potential for identifying voting patterns, though there remains room for improvement in refining its predictive capabilities by incorporating additional factors such as demographic shifts, voter turnout variability, and emerging political dynamics.

Figure 4: 2024 Presidential Election Predicted Results Map

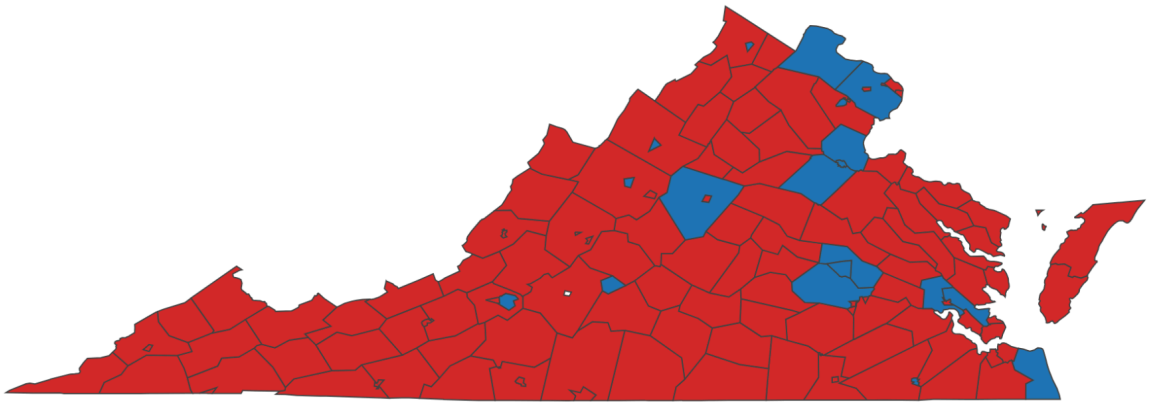

Figure 5: 2024 Presidential Election Actual Results Map

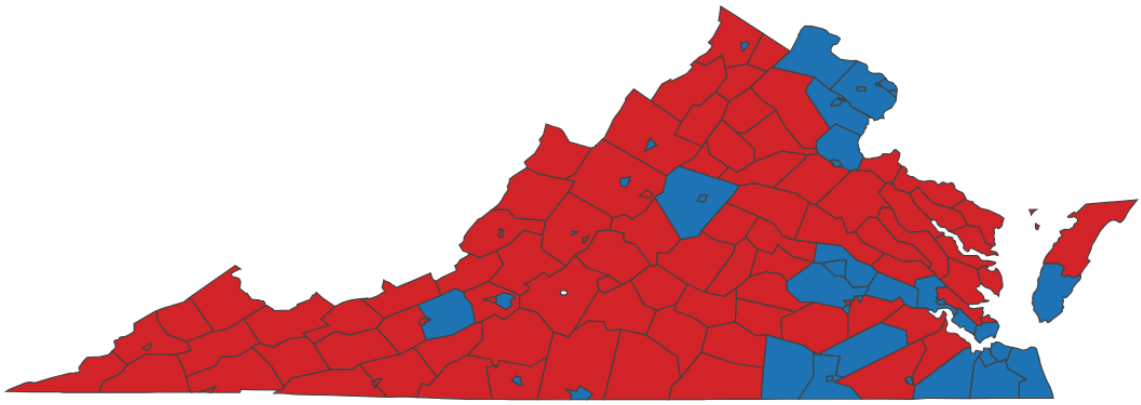

Figure 4 shows the 2024 presidential election results that our logistic regression model predicted. The map in Figure 4 shows a strong prediction of Republican dominance (red) across most counties, with only a few urban or suburban counties predicted to vote Democratic (blue). Figure 5 shows the actual 2024 presidential election results. In the actual results map some counties predicted to vote Republican appear blue, indicating a shift in party support. There is also a noticeable increase in Democratic representation, particularly in regions where urban and suburban areas are located. Generally, we were correct in predicting the Greater Richmond area and Northern Virginia counties. We were incorrect in predicting the coastal region outside of Virginia Beach, such as the Eastern Shore and some areas along the James River. These locations voted Democratic rather than Republican as we predicted. The same is true for Brunswick, Greensville, and Sussex counties.

V. Results Conclusion

Our analysis demonstrates the utility of historical voting trends in predicting election outcomes, while acknowledging limitations of the model. Using a logistic regression model, we achieved a 78% accuracy rate in forecasting the results of the 2024 presidential election for Virginia counties by leveraging features such as party vote share trends, voter turnout changes, and margins of victory. Overall, the model did well in predicting Republican outcomes but struggled slightly with Democratic wins. Comparing our predictions to actual results revealed areas of strength, such as correctly forecasting outcomes in Northern Virginia and Greater Richmond, and areas for improvement, particularly in regions like the Eastern Shore and the James River area. These discrepancies underscore the importance of accounting for external variables, demographic shifts, and emerging political dynamics in future iterations of our model.

CONCLUSION:

This study aimed to evaluate how historical voting data and the trends we can derive influence the ability to predict the outcomes of U.S. presidential elections in Virginia counties. By running this regression on the election data from 2000 to 2020, with the target prediction variable being the election outcome (1 for Democrat and 0 for Republican), it resulted in a 77% accuracy for training data and 74% for testing data. Our next step was to use our logistic regression model to predict the outcome of the 2024 election for each county in Virginia. The predicted outcomes are mapped in Figure 4 above. When comparing these outcomes to the actual results in Virginia (*2024 Virginia Election Results*, 2024), we discovered an accuracy of 78%.

The features that most impacted the model were the trend of Democratic vote share, with a coefficient of -0.35, and the trend of Republican vote share, with a coefficient of -3.48. We

want to keep in mind that our model classified a Democratic win as a 1, so these coefficients are in relation to the likelihood of a Democratic win. The large negative coefficient for the Republican share trend makes sense in this context, as it implies that as the percentage of votes the Republican party receives increases over the years, the likelihood of a Democratic outcome in that county greatly decreases. However, we found the negative coefficient for the Democratic share trend interesting, since it signifies that as the percentage of votes the Democratic party receives increases over the years, the likelihood of a Democratic outcome decreases. Though it seems counterintuitive, an increasing trend in Democratic votes does not correlate directly to a Democratic win, as a majority of counties in Virginia are historically Republican dominated. Thus, we can determine that many counties are areas where Republican gains may outpace Democratic trends, even if there is an emerging increase. This negative coefficient also suggests that counties with a higher rate of Democratic vote share growth may still be influenced by other factors when determining the overall winner.

The other factors played a vital and nuanced role in predicting election outcomes in Virginia counties, though their coefficients were less impactful (Figure 3). The coefficient for trend in votes all other political parties receive is negative and slightly lower than the trend of the Democratic percentage. This implies that as other parties grow in popularity, this takes away from votes the Democratic party could be receiving, thus reducing the likelihood of the Democratic party coming out on top. Similarly, we can see that as the margin of victory increases between the two top parties, these being Democratic and Republican for all the records in our dataset, it is more likely that a Republican will win. We can interpret this as saying that when there is a party that takes a greater majority of the votes, this tends to be the Republican party. Finally, as voter turnout increases between elections, it is more likely for that election to lean

favorably towards a Democratic candidate. All of these findings create a picture of voting behavior in Virginia, revealing that many counties are predominantly Republican and that the fluctuations in this party specifically are a strong indicator of how the citizens of that county will vote.

Our logistic regression model provides a framework for understanding county-level dynamics and can assist political strategists in targeting resources and tailoring campaign messages. It acts as a quantitative tool to predict the outcome in future elections by inputting projected values for feature variables. Thus, the model could be particularly useful for political analysts and academics during election seasons, since creating an idea of how a state will vote is an important task during the time leading up to an election. In particular, our model takes voting behavior into account and can be extended to evaluate long-term trends in different counties. These trends also provide key indicators for counties that are vulnerable to low voter turnout and provide a basis for academics to target initiatives to improve voter education and engagement. By analyzing the change in voter turnout as a feature variable, campaigns may focus efforts on voter mobilization in key counties. Additionally, predictive models like this can also help identify swing counties where changes in voter turnout or party support trends could significantly impact election results. If there are counties where there is a significant positive trend in Republican vote share or a negative Democratic vote share, this could indicate swing counties and areas where Democratic campaigns might need to increase voter outreach to counteract Republican gains.

The model's reliance on Republican voting trends and its tendency to underestimate Democratic wins may stem from the absence of demographic and population variables. The current data, which heavily reflects Republican voting patterns, likely biases predictions toward

Republican wins, in part due to the lack of detailed data on Virginia county populations. Incorporating more comprehensive population data is essential for accurately representing voting trends in Virginia. Future research should integrate demographic factors, population density, and alternative classification methods to enhance predictive accuracy. The omission of key variables such as these can lead to a model that underfits and adjusts to compensate for missing information. Additionally, examining interactions between variables, such as the relationship between turnout changes and normalized margins, could provide deeper insights into voting behavior. The goal of our logistic regression model is the prediction of future presidential elections, which leads us to accept a certain level of omitted variable bias in order to create more stable and reliable predictions.

Understanding the drivers of election outcomes is essential for fostering an informed electorate and strengthening democratic processes. This study contributes to the effort by identifying quantifiable factors that influence voting patterns in Virginia counties. Furthermore, our approach demonstrates how machine learning can be applied to understand the nuanced behavior of political science. The methodology serves as a blueprint for developing predictive models in other states or at other levels of government, making it a valuable tool for studying electoral trends nationwide. By emphasizing the significance of treds, turnout, and margins of victory, we highlight areas for action to strengthen the democratic process. Ultimately, our findings contribute to a deeper understanding of the challenges and opportunities facing elections and utilize machine learning in order to make data-driven, democratic decisions.

REFERENCES:

2024 Virginia Election Results. (2024). WVTF. https://www.wvtf.org/2024-virginia-election-results.

Plotly. Plotly Python Graphing Library. (n.d.). https://plotly.com/python/